Cascade Region Regression for Robust Object Detection

Jiankang Deng
Team Name: Amax

Centre for Quantum Computation & Intelligent Systems (QCIS), University of Technology Sydney (UTS)
Jiangsu Key Laboratory of Big Data Analysis Technology (B-DAT), Nanjing University of Information Science & Technology (NUIST)
Submission Brief
(With Additional Training Data)

- **Object detection (DET)**
  rank 1# (Categories won: 165/200, mAP: 0.57848)

- **Object localization (LOC)**
  rank 2# (Loc error: 0.14574, Cls error: 0.04354)

- **Object detection from video (VID)**
  rank 1# (Categories won: 18/30, mAP: 0.730746)

Key idea: **Cascade Region Regression**
  “Where” from a former layer, and “What” from a later layer
  Answering “where” more accurately helps answer “what”
R-CNN

General framework: Region proposal + DCNN based region classification

1. **Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition**, Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, in ECCV 2014

2. **Object Detection Networks on Convolutional Feature Maps**, Shaoqing Ren, Kaiming He, Ross Girshick, Xiangyu Zhang, Jian Sun, in arXiv 2015

Improving R-CNN

Receptive Field: 171 and 228 pixels for ZF and VGG.

Observations:
1. More accurate and less number of proposal boxes improve the region classification performance. (Fast R-CNN vs Faster R-CNN)
2. High capacity model usually leads to high performance. (ZF vs VGG)

Question:
Location indexed features are able to regress more accurate boxes. What’s the condition? 0.7IoU? 0.5IoU? 0.4IoU?

Our Method

Diagnosis experiments on val2
Faster R-CNN Baseline

Training procedure:
1. Train Faster R-CNN on ILSVRC2014_train and Validation1.
2. Get the scores of the annotation boxes on all training data.
3. Remove the incorrect annotations (associated with low scores).
4. Add omissive annotations (associated with high scores).
5. Test the model on ILSVRC2013_train data set.
6. Remove easy training data (large single object).
7. Train Faster R-CNN on the refined training data.
Easiest and hardest categories

- Large object area within box
- Discriminative appearance or shape
- Small variance
- More training data

- Very small object area within box
- Thin objects
False Positive examples

The box is too small.

The box is too large.

The box covers dense objects.

Inaccurate localization induces false positives.
False Positive examples

Classification error induces false positives.
False Positive Analysis

NoC (region based training)

- Cor: 69.6%
- Loc: 20.1%
- Sim: 1.6%
- Oth: 1.3%
- BG: 7.4%

Fast R-CNN (image based training)

- Correct: 71.6%
- Background: 13.6%
- Sim: 4.3%
- Other: 1.9%
- Loc: 8.6%
Cascade Region Regression

Multi-layer Conv Feature
(region size specific)

Multi-scale Conv Feature
(object + around context)
Conditions of Initial location

Class-wise energy / box receptive field energy is highly related to the probability of convergence.

In practice, we define positive examples which can regress better locations (or keep).

Learning to Combine

Object detection via a multi-region & semantic segmentation-aware CNN model, Spyros Gidaris, Nikos Komodakis, in ICCV 2015
Learning to rank

Class-specific classifier is trained by SPP-net (multi-scale).

Suppress false positives from background.
## Additional Training Data

<table>
<thead>
<tr>
<th>Class Name</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>accordion</td>
<td>4.27%</td>
</tr>
<tr>
<td>ant</td>
<td>5.64%</td>
</tr>
<tr>
<td>armadillo</td>
<td>3.93%</td>
</tr>
<tr>
<td>balance beam</td>
<td>7.33%</td>
</tr>
<tr>
<td>banjo</td>
<td>15.46%</td>
</tr>
<tr>
<td>baseball</td>
<td>4.05%</td>
</tr>
<tr>
<td>bee</td>
<td>4.72%</td>
</tr>
<tr>
<td>binder</td>
<td>2.32%</td>
</tr>
<tr>
<td>bow tie</td>
<td>3.54%</td>
</tr>
<tr>
<td>bow</td>
<td>3.63%</td>
</tr>
<tr>
<td>......</td>
<td>......</td>
</tr>
</tbody>
</table>

### Detection (thre=0.5)
- Remove FP, Add FN, Refine boxes

Add training data.
Model ensemble is always effective.
Diagnosis experiments on val2
Object detection from Video

Object detection on each frame

Tracking from the high score frame (temporal smooth)

Class-wise box regression and NMS on each frame
Object detection from Video

Scene Cluster (object detection + similarity scene)
Scene Context is helpful to suppress FP.
Team member

Jiankang Deng  Shaoli Huang  Jing Yang  Hui Shuai  Zhengbo Yu  Zongguang Lu

Qiang Ma  Yali Du  Yi Wu  Qingshan Liu  Dacheng Tao